Библиографические ссылки

1. Новости про операционные системы. 06.08.2013 [Электронный ресурс]. Режим доступа: http://nvworld.ru/news/android-controling-87-percent-tablet-market/

2. Digit : интернет-журнал о технологиях. 23.08.2013 [Электронный ресурс]. Режим доступа: http://digit.ru/technology/20130823/404655676.html

ИДЕНТИФИКАЦИЯ АВТОРА ТЕКСТА ПО СТОХАСТИЧЕСКИМ ХАРАКТЕРИСТИКАМ ПИСЬМЕННОЙ РЕЧИ

Н. Н. Бороденко
(Екатеринбург, УрФУ, natalia@gammaural.ru)

Вопрос об идентификации автора текста в глобальной сети стал одним из наиболее обсуждаемых представителями государственных органов и обществом.

Так, в Российской Федерации предложения о законодательном «запрете анонимности» в Интернете неоднократно высказывались руководителями правоохранительных ведомств в контексте борьбы с преступностью.

Правомерно пытаться найти ответы на следующие ключевые вопросы, связанные с идентификацией в Интернете:

1) можно ли создать универсальную, всеобщую, глобально признанную систему идентификации пользователей Интернета.

2) если создание такой системы возможно, то на каких принципах и с использованием каких технологий. Каковы могут быть цели такой идентификации. Как избежать в процессе ее создания и использования нарушений основных прав и свобод человека, в том числе на неприкосновенность частной жизни.

3) если создание системы, упомянутой в предыдущем пункте, невозможно, то по каким основным технологическим, организационным, правовым либо иным причинам.

1. Основные методы идентификации пользователя в сети

Существуют следующие основные методы идентификации пользователя сети:

1) сопоставление профилей из социальной сети (это задача разрешение объекта) и общая идея заключается в следующем: есть два сайта с профилями, предлагаемая автором модель выявляет, какие профили первого сайта соответствуют профилям второго сайта, посредством чего возможно предположить, что человек имеет не более одного профиля на каждом сайте [1];

2) история запросов в браузере, ведь большинство из нас (около 70 %) с точностью до 97 % случаев точно определяется по факту и частоте просмотра всего 4 (уникальных для каждого из нас) сайтов;

3) основные методы идентификации пользователей: по IP-адресу, cookie, сессионным идентификаторам, авторизации пользователей.

Метод идентификации автора в сети посредствам изучения последовательности букв (биграмм и триграмм) в текстах пользователя является принципиально новым.

2. Описание метода

Предполагается, что в пространстве Интернета на одном из сайтов (в качестве примера возьмем сайт http://www.proza.ru/ (s_1)), пользователь однозначно идентифицировал себя (ввел фамилию, имя и отчество). Присутствующие на сайте тексты принадлежат именно ему. На исследуемом сайте (http://soyuz-pisatelei.ru/ (s_2)) однозначная идентификация отсутствует (пользователи сайта «Союз писателей» идентифицируются по придуманному в процессе регистрации логину). Предлагаемый метод будет сравнивать элементы текстов с сайтов s_1 и s_2, в результате чего формировать вывод о принадлежности текста с сайта s_2 идентифицированному пользователю.

Кроме того, метод позволит в некоторой степени решить проблему умышленного присвоения авторства чужого произведения (плагиата) и из достаточного по объему текста определить количество авторов одного текста.

Стоит подчеркнуть, что анализируется только последовательность букв того алфавита, на котором написан текст.

Анализируемой текстовой структурой принимаем решение считать плотность вероятности текста по буквосочетаниям или n-ПВ, где n отвечает за порядок n-грамм.

Следует отметить, что строго математического доказательства этой идеи быть не может, так как автор вправе написать произведение в любой, даже не свойственной ему, манере. Для этого достаточно наложить на автора некоторые ограничительные рамки: стиля, темы, синтаксиса [2].

3. Суть метода

Обозначим задачу идентификации автора неизвестного текста: имеется библиотека, содержащая тексты, представленные в виде ПВ для a известных авторов. Пусть T_a – имеющееся количество текстов a-го автора, и $N_{i,a}$ – количество букв в i-м тексте этого автора, $i = 1, 2, \ldots, T_a$.

В частности за a-го автора примем Испытуемого № 1. Имеется 29 текстов данного автора, опубликованные на сайте http://www.proza.ru/ (длина каждого текста достаточна для проведения статистического анализа и составляет не менее 2000 букв). 6-й текст (миниатюра «Ошибочка») состоит из 361 слова и 2888 букв (рис. 1, 2).

Примем f_i, $a(j)$ – n-ю ПВ соответствующего текста, где аргумент j меняется от 1 до $\alpha(n) = 33n$.

Тогда для Испытуемого № 1 средневзвешенная ПВ:

$$F_a(j) = \frac{1}{N_a}\sum_{i=1}^{T_a} f_{i,a}(j)N_{i,a},\ N_a = \sum_{i=1}^{T_a} N_{i,a}. \qquad (1)$$

Эта ПВ (1) будет играть в дальнейшем роль авторского эталона (автор – Испытуемый № 1).

Введем $R_{i,k}$ (2) как расстояние между ПВ текстов i и k (i – «Ошибочка»; k – «Опасная встреча»):

$$R_{j,k} = |f_i - f_k| = \sum_{j=1}^{\alpha(n)} |f_i(j) - f_k(j)|. \qquad (2)$$

В процессе исследования определим величину $(1 - G_a^+(R_a^-))$ – вероятность ошибочно признать за произведение автора a чужой

Рис. 1. Распределение расстояний между 3-ПВ текста и авторского эталона

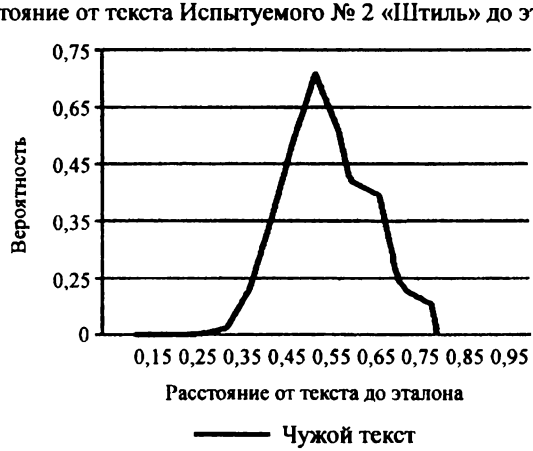

Рис. 2. Распределение расстояний между 3-ПВ текста и авторского эталона

текст (ошибка второго рода), а величину $G_a^-(R_a^+)$ – вероятность ошибочно отвергнуть произведение автора a, приняв его за чужое (ошибка первого рода) [3].

Значение $\hat{R}$ (3) будем считать расстоянием разделения авторов, для которого ошибка идентификации автора минимальна [4]:

$$\hat{R} = \arg_{\min}(1 - G^{+}(R) + G^{-}(R)) = \arg_{\max}(G^{+}(R) - G^{-}(R)). \quad (3)$$

Естественно, на момент сравнения произведение с установленным авторством исключается из эталона (1), так что ПВ Испытуемого № 1 сравнивалась с его квазиэталоном:

$$F_{i,a}(j) = \frac{1}{1 - \frac{N_{i,a}}{N_a}} (F_a(j) - f_{i,a}(j) \frac{N_{i,a}}{N_a}, \quad (4)$$

$$\left| f_{i,a} - F_{i,a} \right| = \frac{1}{1 - N_{i,a} / N_a} \sum_{j=1}^{\alpha(n)} \left| f_{i,a}(j) - F_a(j) \right| = \frac{\left| f_{i,a} - F_a \right|}{1 - N_{i,a} / N_a}. \quad (5)$$

Пусть имеется текст «Столкновение» неизвестного автора, который надо идентифицировать внутри данной библиотеки. Автором текста «Столкновение» считается тот из авторов, для которого норма $R_a^1 = |f_0 - F_a|$ разности между ПВ $f_0(j)$ текста «Столкновение» и средней авторской ПВ $F_a(j)$ минимальна:

$$a^1 = \arg_{\min} R_a^1. \quad (6)$$

Правило (6) принимается только в том случае, если $\min R_a^0 \le \hat{R}$. Если же оказалось, что $\min R_a^0 > \hat{R}$, т. е. минимальное расстояние превосходит длину разделения, то принимается решение об отсутствии в библиотеке подходящего авторского эталона [3].

В ходе исследования проанализировали некоторые характеристики расстояний между текстом и эталоном, которые приведены в таблице.

В ходе исследования были приведены аргументы в пользу эффективности метода 3-ПВ по сравнению с распределениями другой размерности. Во-первых, ошибка второго рода, являющаяся в задаче идентификации автора ключевой, минимальна для 3-ПВ. Во-вторых, для распределений расстояний от текстов до «своих–

Характеристики расстояний между текстом и эталоном

Показатель	1-ПВ	2-ПВ	3-ПВ
Среднее значение l_s (свой текст)	0,013	0,055	0,032
Среднее значение l_d (чужой текст)	0,024	0,084	0,083
Стандартное отклонение σ_s (свой текст)	0,057	0,103	0,057
Стандартное отклонение σ_d (чужой текст)	0,079	0,113	0,059
Расстояние разделения $\hat{R}$	0,043	0,187	0,342
R^+	0,106	0,291	0,561
R^-	0,033	0,120	0,261
Вероятность ошибки I рода $G^-(R^+)$	0,469	0,626	0,221
Вероятность ошибки II рода $1 - G^+(R^-)$	0,723	0,529	0,301

чужих» эталонов площадь перекрытия графиков также минимальна для 3-ПВ. В третьих, если вычислить разность средних расстояний от текста до «чужого» и до «своего» авторов $l_d - l_s$ и сравнить ее с суммой соответствующих дисперсий $\sigma_d - \sigma_s$, то для 1- и 2-ПВ меньше, чем для 3-ПВ, т. е. различающая способность авторских эталонов здесь наибольшая (вероятность правильного обнаружения 86–91 %).

Идентификация текстов может проводиться с большей надежностью, если у автора достаточно произведений, написанных в разных жанрах, и тогда можно было бы рассмотреть вместо одного несколько авторских эталонов.

Для авторов, чьи тексты состоят не менее чем из 2000 букв, исследуемый метод показал высокоточный результат (проанализировано более 200 текстов писателей и поэтов в Интернет-пространстве).

Однако на практике чаще возникает задача идентификации автора, не являющегося в прямом смысле профессиональным писателем. Тогда возникает вопрос о необходимой и достаточной длине текста. Так как в жизни иногда возникает потребность провести анализ не большого произведения, а выражения (в частности динамического), состоящего из 5–10 слов, то здесь вместе с букво-

сочетаниями должны исследоваться и другие критерии оценки текста для увеличения вероятности правильного установления авторства (каковы они – это предстоит выяснить в ходе дальнейших исследований).

Библиографические ссылки

1. Maurice van Keulen Matching Profiles from Social Network Sites, 2009 [Электронный ресурс]. URL:http://wwwhome.cs.utwente.nl/~keulen/wordpress/2009/10/matching-profiles-from-social-network-sites/

2. *Андреев Н. Д.* Статистико-комбинаторные методы в теоретическом и прикладном языковедение. М. : Наука, 1967. 402 с.

3. *Валгина Н. С.* Теория текста : учеб. пособие. М. : Логос, 2003. 280 с.

4. *Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф.* Справочник по теории вероятности и математической статистике. М. : Наука, 1985. 640 с.

ФОРМИРОВАНИЕ БАЗЫ РЕШАЮЩИХ ПРАВИЛ СИСТЕМЫ ОБНАРУЖЕНИЯ АТАК С ПОМОЩЬЮ ГЕНЕТИЧЕСКОГО АЛГОРИТМА

А. О. Власов
(Екатеринбург, УрФУ, ale_vlas@mail.ru)

Формирование базы решающих правил для системы обнаружения атак является одной из главных задач при защите от компьютерных атак. Качественно построенная система правил позволяет выявлять и предотвращать опасные или потенциально опасные воздействия на автоматизированные информационные системы.

Для пользователей и администраторов готовых к использованию «коробочных» систем обнаружения атак формирование и разработка правил распознавания не являются важной задачей: производители зачастую не включают в функционал подобных систем использование пользовательских правил, а имеющиеся базы регулярно обновляются с серверов разработчиков. Задача формирования базы правил в первую очередь интересна производителям сис-